PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C07K 1/00, G06F 17/50

A1

(11) International Publication Number: WO 98/18814

(43) International Publication Date: 7 May 1998 (07.05.98)

(21) International Application Number: PCT/US97/19673

(22) International Filing Date: 27 October 1997 (27.10.97)

(30) Priority Data:

60/029,521	28 October 1996 (28.10.96)	US
60/037,281	3 February 1997 (03.02.97)	US
Not furnished	22 October 1997 (22.10.97)	US

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published

With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(71) Applicant (*for all designated States except US*): CUMULATIVE INQUIRY, INC. [US/US]; 71 Winslow Avenue, Somerville, MA 02144 (US).

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): HALITSKY, David [US/US]; 2619 Waltham Drive, Huntsville, AL 35898 (US). FRESCO, Jacques, R. [US/US]; 282 Hartley Avenue, Princeton, NJ 08544 (US).

(74) Agents: MYERS, Paul, Louis et al.; Lahive & Cockfield, LLP, 28 State Street, Boston, MA 02109 (US).

(54) Title: NUCLEIC ACID-LEVEL ANALYSIS OF PROTEIN STRUCTURE

(57) Abstract

Methods for analyzing protein structure are disclosed. The methods of the invention permit identification of polypeptides which have structural homology to a known polypeptide, but have little sequence homology to the known polypeptide. The methods of the invention are useful for designing novel proteins having desired structural or functional characteristics.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

NUCLEIC ACID-LEVEL ANALYSIS OF PROTEIN STRUCTURE

Background of the Invention

The invention relates to methods of evaluating, altering, and designing protein structures.

Summary of the Invention

Methods of the invention incorporate considerations of mRNA sequence and structure and codon-anticodon energetics into the analysis and design of protein structure. Many prior art methods for analyzing or designing protein structure have relied in part or in whole on analysis at the amino acid level. Proteins, however, are the product of a process which involves a number of cellular entities and their interactions. The interaction of mRNA molecules with the protein translation machinery (e.g., ribosomes, and tRNAs, as well as other elements of the cellular environment, such as water and salt molecules) and mRNA intrachain interactions place physico-chemical restraints on the overall process.

Not only are proteins the product of a process, but the process itself has evolved over time. Some constraints, e.g., those imposed by the interaction of mRNAs with environmental elements or with primitive ribosomal structures, those of mRNA structure and energetics, e.g., the propensity to form secondary structure, may have been more important, or at least different, primordially than they are currently. While not wishing to be bound by theory, the inventors postulate that evidence of those prior constraints may be seen in the sequence of current messages.

Methods of the invention provide for the analysis and design of protein structures on the basis of patterns or features of the nucleic acid message, e.g., codon usage patterns or coding modalities. Methods of the invention are based on dividing the genetic code, that is the codon-anticodon pairs which specify amino acids (and stops), into classes, sometimes referred to herein as subcodes or coding modalities, and evaluating a nucleic acid sequence which encodes a protein structure based on its class (i.e., subcode or coding modality). Relevant subcodes or coding modalities can be defined using choice parameters which are a function of message-level properties, wherein each property is related to the composition or structure of the nucleic acid, and is other than the identity of the amino acid (or stop) encoded and other than codon bias. Examples of structural choice parameters, which can serve as methods or rules for assignment of codons into classes, include the nature of the substituents on the coding

bases (e.g., so-called keto-rich bases U and G or amino-rich bases A and C), size of the coding bases (e.g., purine vs. pyrimidine), hydrogen-bonding and base-stacking energies of the coding bases in overlapping base pairs, and the like. Examples of compositional choice parameters include frequencies of subclasses of codons within more than one of the three alternative reading frames in which a nucleic acid message can be read. Alternative subcodes or coding modalities are not necessarily entirely disjointed, discrete, or unique, and identical subcodes or coding modalities can be obtained using structural and/or compositional parameters.

Methods of the invention allow the identification, analysis, modification and design of protein structures on the basis of patterns or features revealed by the nucleic acid, e.g., the messenger nucleic acid. For example, the identification of a "run" of amino acids residues of a class can be indicative of an evolutionarily conserved region. The identification of a "minority" class codon in a run of majority class codons can be indicative of a structure- or function-critical residue. The discovery of a critical residue can be used in the design or modification of a protein, e.g., to develop a second generation protein. For example, in situations where it is desirable to alter structure or activity of a protein, it may be desirable to alter a critical residue(s) (or a residue which interacts with a critical residue, e.g., an adjacent residue or a residue elsewhere in the protein (or in another protein) with which it interacts). In the case where a change which does not result in significant alterations in structure or activity is desired, residues other than the identified critical residue (or other than residues which interact with it) are changed.

Methods of the invention provide for nearest neighbor frequencies calculated based upon the frequency or pattern of selected classes of codons, i.e., by codon class of the amino acid, and thus provide a higher degree of relevance for analysis of single-class-rich protein structures. Conventional tables of nearest neighbor amino acids do not take into account the classes described herein, and as such, provide only "average" values across multiple classes of codons. Also, unlike tables of the invention, conventional nearest-neighbor tables do not take into account the fact that consistent secondary/tertiary structures of proteins can be shown to correlate with: a) "out of frame" properties of protein messages; b) "interframe" properties of protein messages, i.e., correlations between properties of messages read in frame 1, properties of messages read in frame 2, and/or properties of messages read in frame 3, as defined below.

In general, the invention features a method of evaluating protein structure. The method includes:

providing a nucleic acid sequence which encodes the protein structure;

assorting bases of the nucleic acid sequence into subject triplets; and

assigning one or a plurality of subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying triplets of the nucleic acid sequence as members of a class of a binary choice alphabet of n degrees of freedom, and wherein the classes can be generated by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein a binary choice parameter is a function of a message-level property of the nucleic acid sequence,

thereby evaluating the protein structure. Triplets can be assigned to a class based on whether they satisfy a value for the message level property, e.g., a triplet can be assigned to a class based on whether its value for a parameter is above or below a predefined value, e.g., enthalpy for formation of a codon-anticodon duplex, or whether or not it possess a particular characteristic, e.g., whether it is GC rich. The message-level property is other than, the identity of the amino acid or punctuation which a triplet encodes and is other than codon bias.

The class constant table provides a measure of the frequency with which a first and a second amino acid occur as nearest neighbors and wherein nearest neighbor frequencies are determined within a codon class, and wherein a class is a function of a message level property of a nucleic acid, e.g., the codon, which encodes an amino acid. The class can be any class generated by the binary choice parameter-based methods referred to herein. For example, if the classes are a first class, e.g., high enthalpy codons and a second class, e.g., low enthalpy codons, the table is generated for nearest neighbors where both neighbors are encoded by codons of either the first class or codons of the second class.

In another aspect, the invention features a method of evaluating a protein structure. The method includes: providing a class-constant table of nearest neighbor relationships for amino acid residues;

providing a nucleic acid which encodes a protein structure; and

comparing one or a plurality of the observed nearest neighbor pairs in the protein structure with the frequencies provided by the class constant table, thereby evaluating the protein structure.

In preferred embodiments, the comparison can include: assigning an expected frequency from the class constant table to one or a plurality of the observed nearest neighbor pairs and determining how many of the observed nearest neighbor pairs fall above or below a predetermined value; determining the likelihood of occurrence, as predicted by the class constant table, for an observed nearest neighbor pair.; or determining if an observed nearest neighbor pair of a first and a second amino acid residue from the protein structure is predicted by the class constant table to occur at a predetermined frequency.

In another aspect, the invention features a method of evaluating a protein structure for resistance to change, e.g., evolutionary or mutational change. The method includes:

identifying regions of a protein which is encoded by runs of a single subcode, thereby identifying regions which have been resistant to change and which are therefor predicted to be functionally or structurally significant. E.g., the method can include determining if the nucleic acid sequence which encodes the protein structure includes a run of triplets, e.g., a run at least 20, 40, 60, or 120 triplets in length, in which at least 20, 40, 60, 80, 90 or 95 %, or all, of the triplets in the run are from one class. Any of the ways of generating classes described herein can be used in this method.

In another aspect, the invention includes, a method of evaluating a protein structure for the presence of critical amino acid residues. The method includes:

identifying critical amino acid residues by identifying "minority codons" in runs encoded by codons of a single class or subcode, thereby identifying residues which have been resistant to change and which are therefor believed to be functionally important. Any of the ways of generating classes described herein can be used in this method.

In another aspect, the invention features a method for evaluating a protein structure. The method includes:

providing a nucleic acid sequence which encodes the protein structure;

assorting bases of the nucleic acid sequence into subject triplets; and

assigning at least one of the subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets, the at least four classes of triplets

being represented in at least a portion of the nucleic acid sequence in a ratio of about 3:5:3:5;

thereby evaluating the protein structure.

In another aspect, the invention features a method for identifying coding regions of a nucleic acid sequence, the method comprising:

providing the nucleic acid sequence;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets;

assigning the plurality of subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets A, B, C, and D;

determining whether the plurality of subject triplets are distributed into the at least four classes of triplets A:B:C:D in a ratio of about 3:5:3:5;

thereby identifying coding regions of the nucleic acid sequence.

In another aspect, the invention features, a method for identifying a protein that includes a polypeptide portion which is structurally or functionally similar to all or a portion of a test protein, the method comprising:

providing a nucleic acid sequence which encodes all or a portion of the test protein;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets in a first reading frame;

assigning the plurality of subject triplets in the first reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a first binary choice alphabet of n degrees of freedom by applying n first binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

assorting bases of the at least a portion of the nucleic acid sequence into a plurality of subject triplets in a second reading frame;

assigning the plurality of subject triplets in the second reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a second binary choice alphabet of n degrees of freedom by applying n second binary choice parameters to a triplet to yield at least 2^n

classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5; and

identifying a protein which includes a polypeptide portion encoded by the plurality of triplets in the second reading frame;

thereby identifying a protein that includes a polypeptide portion which is structurally or functionally similar to all or a portion of the test protein.

In another aspect, the invention features, a method for identifying a mutation-prone region of a nucleic acid sequence, e.g., a viral nucleic acid sequence. The method includes:

providing the nucleic acid sequence;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets in a first reading frame;

assigning the plurality of subject triplets in the first reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

assorting bases of the at least a portion of the nucleic acid sequence into a plurality of subject triplets in a second reading frame; and

assigning the plurality of subject triplets in the second reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

thereby identifying a mutation-prone region of the nucleic acid sequence.

In another aspect, the invention includes, a method of providing a protein structure, e.g., the structure of a protein of known function, in which one or a plurality of amino acid residues are changed. The method includes:

providing a nucleic acid sequence which encodes a candidate protein structure;

evaluating the sequence by a method described herein; and

altering one or a plurality of amino acid residues in the candidate protein structure,

thereby providing a protein structure.

In yet another aspect, the invention features, a machine-readable data storage medium, including a data storage material encoded with machine readable data which, when used with a machine programmed with instructions for using the data, is capable of storing, retrieving, or displaying databases, binary choice alphabets, protein sequences, nucleic acid sequences of the invention. The storage medium can be used in methods of the invention. In preferred embodiments the storage medium is recorded with: a class constant nearest neighbor table; the classes into which the triplets of a nucleic acid are assigned; or nucleic acid sequence which encodes or protein structure which is to be analyzed or which has been altered by application of a method described herein.

Methods referred to herein can further include creating a record of one or more protein structures to be analyzed or modified, e.g., proteins, protein portions or fragments, or nucleic acids which encode all or part of such protein structure. The protein or nucleic acid structure which is to be analyzed or modified, or the structure which has been identified, evaluated or modified, or both, can be recorded. The record can be encoded in the form of a machine-readable data storage medium. The recorded structure, e.g., a nucleic acid or amino acid sequence, can be displayed on a machine, e.g., on a monitor, or in printed form.

Methods referred to herein can further include providing an identified or modified substance, e.g., a protein or nucleic acid, e.g., chemically synthesizing the identified substance based on the structure identified by way of the methods described herein. In preferred embodiments, the method includes assessing the biological activity of the identified substance. The biological activity of the identified substance can be assessed *in vitro* or *in vivo*. In preferred embodiments, the identified substance can be combined with a carrier suitable for introduction into any living cell or organism, e.g., an animal model, e.g., naturally derived or synthetic polymers, solvents, dispersion media, coatings, antibacterial and antifungal agents and the like.

Methods referred to herein can further include providing a three dimensional representation of the protein structure, or a representation of the primary sequence of the protein structure, either before or after a modification. The structure can be compared to the candidate structure or can be evaluated for the ability to exhibit a predetermined structure, e.g., possession of a structural component such as a helix, or a turn segment, an activity, e.g., the ability to dock with a second protein.

In methods referred to herein the nucleic acid sequence can be any of: a genomic sequence; an mRNA sequence; a sequence which encodes a protein structure of known function; a sequence for which the reading frame, if it exists, is known; a sequence for which the reading frame, if it exists, is unknown; a sequence which includes a coding portion; a sequence which includes a non-coding portion; or a sequence from a multiprotein data base.

Methods of the invention allow a wide variety of information to be extracted from nucleic acid sequences and allow a wide variety of useful manipulations, e.g., the identification of useful protein structures and the design of improved or altered function protein structures. These include, but are not limited to:

providing a protein structure encoded by codons of a first subcode which has a predetermined property of a protein structure encoded by codons of a second class or subcode. This allows: provision of a protein structure having a novel amino acid sequence but which has a desired property, e.g., secondary structure, of a known protein; provision of protein structure with improved or altered function;

identifying regions of proteins which are encoded by runs of codons of a single class or subcode, thereby identifying regions which have been resistant to evolutionary or mutational change and which may therefore be functionally important;

identifying a critical amino acid residue(s) in a protein structure by identifying "minority codons" in runs encoded by codons of a single class or subcode, thereby identifying amino acid replacements which, although disfavored at the mRNA level, exhibit sufficiently favored characteristics at the protein level that they have been maintained and may therefore be functionally important;

determination of nearest neighbor relationships based upon nearest neighbors encoded by codons drawn from the same class or subcode;

distinguishing a coding region from a non-coding region by determining whether the region obeys nearest neighbor relationships involving codons drawn from the same class or subcode;

assignment of function (or structure) to a protein or polypeptide of unknown structure by recognizing codon patterns in message-level nucleic acid which encodes the protein or polypeptide structure of unknown function (e.g., the protein or polypeptide is encoded in a first subcode) similar to codon patterns in message-level nucleic acid which encodes the structure of known function (but different primary sequence) (e.g., which is encoded by a second subcode).

DEFINITIONS

As used herein, "protein structure" refers to a structure of at least two amino acids linked by a peptide bond. A protein structure can include an entire protein, or a part thereof. For example, a protein structure can include a domain or other region having a characteristic structural, chemical, or biological property. Examples of structural elements include helices, turns, sheets, helix-turn structure; tertiary amino acid structure; and the like. Examples of chemical properties include net charge, side chain bulk, side chain charge, acidity, nucleophilicity, hydrophobicity, and the like. Examples of biological properties include catalytic activity, promoter or suppressor activity, ability to bind to or interact with a second molecule such as DNA, RNA, a protein, a metal atom, immunological activity, and the like. Examples of known domains which can be included in protein structures include: zinc fingers, binding regions, and the like. A protein structural element can be from a naturally occurring protein or can be a non-naturally occurring (e.g., a novel) construct. The protein structure can be of a predetermined length. In preferred embodiments it is at least 8, 16, 32, 64 or 128 amino acids in length.

As used herein, a predetermined property is a property other than the sequence of amino acids, and can include one or more of the following: (1) three dimensional structure, e.g., secondary structure, tertiary structure, or quaternary structure; (2) a charge-related property, e.g., due to positively or negatively charged side chain residues, including, but not limited to: the presence of a predetermined charge at a predetermined location in the sequence, the net charge on a protein or polypeptide, and the like; (3) hydrophobicity, e.g., due to the presence of water-insoluble side-chain residues; (4) an activity associated with an intramolecular interaction or an intermolecular interaction. Intermolecular interactions include binding activity, catalytic activity, and the like.

An "amino acid alphabet," as used herein, refers to a group of codons which encode amino acids or stop codons.

As used herein, a "binary choice" amino acid alphabet of n degrees of freedom, refers to an amino acid alphabet which is structured into 2^n subcodes, by the application of binary choices dictated by n choice parameters, and where a choice parameter is a function of nucleic acid sequence and/or codon patterns of the nucleic acid (e.g., an mRNA).

A "binary choice parameter" or "opposition," as used herein, refers to a parameter by which a polynucleotide codon or triplet can be assigned one of two values. The assigned values allow the triplets to be assigned to classes. It will be appreciated that application of more than one non-degenerate binary choice parameter can divide

triplets into more than two classes. The division into classes can be based on a predetermined value. E.g., all triplets with a value less than the predetermined value are in one class and all with values above the predetermined values are in a second class, or all triplets having predetermined characteristic a, e.g., being pyrimidine-rich, are in a first class and all codons being pyrimidine-poor are in a second class.

The term "coding modality," as used herein, refers to a pattern of codon usage in a nucleic acid message, e.g., the frequency that one or more codons appears in a nucleic acid sequence, the relative frequency that one or more codons appears in two or more reading frames of a nucleic acid message, and the like.

A "triplet", as used herein, refers to three contiguous (sequential) nucleic acid residues (e.g., read in the 5'-3' direction along the nucleic acid strand). A triplet can be a codon (e.g., when a coding nucleic acid sequence is read in the coding frame) or can be a non-reading frame triplet or non-coding triplet.

A leading triplet, as used herein, refers to a triplet which is 5' to the most 3' base in the subject triple. Thus, in a sequence 12345, the leading triplet is 123.

A final triplet, as used herein, refers to a triplet which is 3' to the most 5' base in a subject triple. Thus, in a sequence 12345, the final triplet is 345.

A class of triplets, as used herein, refers to all triplets which fall within a particular subgroup of triplets under a selected binary choice alphabet.

A message-level property, as used herein, refers to a property of a nucleic acid (e.g.,. mRNA) of three or more bases in length, which property is other than the identity of or physical or chemical property of an amino acid (or punctuation) encoded by the nucleic acid (wherein such physical and chemical characteristics include, e.g., size, hydrophobicity, hydrophilicity), and is other than codon-bias. Structural message-level properties include physical and energetic properties of the nucleic acid. Examples include: UA-rich triplets vs. CG-rich triplets; UG-rich triplets vs. AC-rich triplets; purine-rich ("R-rich") triplets vs. pyrimidine-rich ("Y-rich") triplets; assigning a plurality of codons in said sequence to (1) either a Y-rich subcode or an R-rich subcode and (2) to either an E-rich (UG-rich) subcode or an M-rich (AC-rich) subcode. Compositional message-level properties include frequencies of particular codon groups in one or more reading frames of a message.

The term "reading frame" is known in the art and refers to a frame for reading, e.g., translating, a nucleic acid message. For example, a sequence of nucleotides 123456789 can be read in three reading frames (e.g., in groups of three nucleotides, each triplet being a codon): Reading Frame 1: 123 456 789; Reading Frame 2: 234 567; or Reading Frame 3: 345 678.

"Evaluating protein structure," as used herein, refers to determining properties of a protein or polypeptide. For example, evaluating protein structure includes: determining the three-dimensional structure of a protein or polypeptide; comparing the three-dimensional structure of a known protein or polypeptide with that of an unknown protein or polypeptide; determining the function of a protein or polypeptide; comparing the function of a known protein or polypeptide with that of an unknown protein or polypeptide; and the like.

Other features and advantages of the invention will be apparent from the following detailed description, and from the claims.

Brief Description of the Drawings

Figure 1 schematically depicts alternate reading frames for a nucleic acid message.

Figure 2 depicts the distinction between "wildcard" and "constant" codon doublets.

Figure 3 shows the 64 codons divided into four groups based on the "wildcard" and "constant" distinction and the leading base of the codon.

Figure 4 shows the frequencies of codons in the groups of Figure 1 in a test mRNA database.

Detailed Description

In general, the invention features, a method of evaluating protein structure. The method includes:

providing a nucleic acid sequence which encodes the protein structure;

assorting bases of the nucleic acid sequence into subject triplets; and

assigning one or a plurality of subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying triplets, e.g., a subject triplet or a leading and following triplet of the subject triplet, of the nucleic acid sequence as members of a class of a binary choice alphabet of n degrees of freedom, and wherein the classes can be generated by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein a binary choice parameter is a function of a message-level property of the nucleic acid sequence,

thereby evaluating the protein structure. Triplets can be assigned to a class based on whether they satisfy a value for the message level property, e.g., a triplet can be assigned to a class based on whether its value for a parameter is above or below a

predefined value, or whether or not it possess a particular characteristic, e.g., whether it is GC rich.

The message-level property: is other than, the identity of the amino acid or punctuation which a triplet encodes and is other than codon bias.

In preferred embodiments the method includes making a record, e.g., on a machine readable medium, of the class assigned to one or more triplets.

In preferred embodiments: n is chosen from the integers 1, 2, 3, and 4.

In preferred embodiments the message-level property is a function of a physical or chemical property of one or more bases of a nucleic acid; is a function of a physical or chemical property which affects the tendency of a nucleic acid to form secondary structure.

In preferred embodiments triplets are assigned to a first and a second class:

the first class having the property that a message made of triplets drawn exclusively from the first class is less likely to form secondary (intrachain) structure than is a message which is made of triplets from both the first class and the second class of triplets, and

the second class having the property that a message made of triplets drawn exclusively from the second class is less likely to form secondary (intrachain) structure than is a message which is made of triplets from both the first class and the second class of triplets.

In preferred embodiments the message-level property is: a function of the UA content of a subject triplet; a function of the GC content of a subject triplet; a function of the size or molecular weight of a triplet; a function of whether the triplet is keto rich or amino rich; a function of whether the triplet is purine rich or pyrimidine rich; a function of a the enthalpy of the interaction between the triplet and a fully or partially complementary nucleic acid.

In preferred embodiments: the binary choice parameter is applied to the subject triplet, e.g., applied to the codon which encodes an amino acid, to place a subject triplet in a class.

In preferred embodiments: the class into which a subject triplet is assigned is a function of:

(1) providing a value for a subject triplet of bases 456, wherein the value is a function of the application of a binary choice parameter to a first set of contiguous bases which includes all or a subset of the bases of the subject triplet, e.g., bases 4 and 5 and of the application a binary choice parameter to a second, different, set of contiguous bases which includes all or a subset of the bases of the subject triplet, e.g., bases 5 and 6; and

(2) assigning a plurality of subject triplets to a first class, and a plurality of triplets to a second class, as a function of subject triplet value.

In preferred embodiments: the class into which a subject triplet is assigned is a function of:

(1) providing a value for a subject triplet of bases 456, wherein the value is a function of the application of a binary choice parameter to a first subset of the bases of the subject triplet, e.g., 4 and 5, and of the application a binary choice parameter to a second, different, subset of the bases of the subject triplet

(2) assigning a plurality of subject triplets to a first class, and a plurality of triplets to a second class, as a function of subject triplet value.

In preferred embodiments: the class into which a subject triplet is assigned is a function of:

(1) providing a value for a subject triplet of bases 456, wherein the value is a function of $(S^1 + S^2)/2$, wherein S^1 a function of the application of a binary choice parameter (e.g., the value for enthalpy of anticodon-codon formation above or below a predetermined value) to a first subset of the bases of the subject triplet, e.g., bases 4 and 5 of the subject triplet, and S^2 is a function of the application of a binary choice parameter to a second, different, subset of the bases of the subject triplet, e.g., bases 5 and 63 of the subject triplet; and

(2) assigning a plurality of subject triplets to a first class, and a plurality of triplets to a second class.

In preferred embodiments: the class into which a subject triplet is assigned is a function of the application of a binary choice parameter to one or both of a leading triplet or a final triplet of the subject triplet.

In preferred embodiments: the class into which a subject triplet is assigned is a function of:

(1) providing a value, e.g., enthalpy, of a triplet of bases 456, wherein the value is a function of $(S^1 + S^2)/2$, wherein S^1 is the value, e.g., enthalpy, of the base pair doublet 45 of the subject triplet, and S^2 is the value, e.g., enthalpy, of the base pair doublet 56 of the subject triplet; and

(2) assigning a plurality of subject triplets to a first class, e.g., a low enthalpy class, and a plurality of triplets to a second class, e.g., a high enthalpy class.

In preferred embodiments: a subject triplet 456 of a nucleic acid sequence of bases 123 456 789 is assigned into a class as a function of:

(1) performing one or more of (i), (ii), and (iii)

(i) applying a binary choice parameter to a leading triplet of 456, e.g., to one or more of triplet 123, 234, or 345, to yield a leading value;

(ii) applying a binary choice parameter to 456, to provide a center value;

(iii) applying a binary choice parameter to a following triplet of 456, e.g., to one or more of triplet 567, 678, or 789, to yield a following value;

(2) assigning one or a plurality of subject triplets 345 into a class based on the values determined in one or more of (1), (3) and (3).

thereby assigning one or a plurality of subject triplets into classes.

In preferred embodiments: the class into which a subject triplet is assigned is a function of the application of a first binary choice parameter to a leading triplet and a second binary choice parameter to a following triplet of a subject triplet.

In preferred embodiments: the evaluation includes determining if the nucleic acid sequence includes a run of triplets, e.g., a run at least 20, 40, 60, or 120 triplets in length, in which at least 20, 40, 60, 80, 90 or 95 %, or all, of the triplets in the run are from a first class. The method allows for evaluating a protein structure for resistance to change, e.g., evolutionary or mutational change, by identifying regions of the protein which structure encoded by a run of a single class or subcode, thereby identifying regions which have been resistant to change and which are therefor predicted to be functionally or structurally significant. In preferred embodiments a codon, preferably within the run, is changed so as to alter the sequence of the encoded amino acid to provide an altered sequence.

In preferred embodiments: the evaluation comprises identifying a triplet from a first class in a run of triplets of a second class, e.g., a run at least 20, 40, or 60 codons in length, in which at least 20, 40, 80, 90 or 95 %, or all, of the codons are from the second

class, thereby identifying the triplet of the first class as encoding a critical residue, e.g., a structure or function critical residue. In a preferred embodiment, a codon is changed so as to alter the amino acid encoded by the critical residue, a residue adjacent to the critical residue, or a residue which interacts with the critical residue, and thereby provide an altered sequence.

In preferred embodiments: the nucleic acid encodes a protein structure of known or unknown function.

In another aspect, the invention features, a class-constant table of nearest neighbor relationships for amino acid residues which provides, for each of a plurality of class constant nearest neighbors, a frequency of occurrence which is a function of the occurrence of the class constant nearest neighbor pair in a collection of protein structures, e.g., a collection of at least 10, 50, 100, or 500 proteins.

The class constant table provides a measure of the frequency with which a first and a second amino acid occur as nearest neighbors and wherein nearest neighbor frequencies are determined within a codon class, and wherein a class is a function of a message level property of a nucleic acid, e.g., the codon, which encodes an amino acid. The class can be any class generated by the binary choice parameter-based methods referred to herein. For example, if the classes are a first class, e.g., high enthalpy codons and a second class, e.g., low enthalpy codons, the table is generated for nearest neighbors where both neighbors are encoded by codons of either the first class or codons of the second class.

In preferred embodiments: the assignment of amino acids into a class is done by assigning a codon which encodes it into a class as a function of classifying triplets, e.g., the subject codon or a leading and following triplet of the subject codon, as a member of a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of triplets, wherein a binary choice parameter is a function of a message-level property of the nucleic acid sequence.

The table can be recorded on a machine readable medium.

In another aspect, the invention features, a method of evaluating a protein structure. The method includes: providing a class-constant table of nearest neighbor relationships for amino acid residues;

providing a nucleic acid which encodes a protein structure; and

comparing one or a plurality of the observed nearest neighbor pairs in the protein structure with the frequencies provided by the class constant table, thereby evaluating the protein structure.

The class constant table provides a measure of the frequency with which a first and a second amino acid occur as nearest neighbors and wherein nearest neighbor frequencies are determined within a codon class, and wherein a class is a function of a message level property of a nucleic acid, e.g., the codon, which encodes an amino acid. The class can be any class generated by the binary choice parameter-based methods referred to herein. For example, if the classes are a first class, e.g., high enthalpy codons and a second class, e.g., low enthalpy codons, the table is generated for nearest neighbors where both neighbors are encoded by codons of either the first class or codons of the second class.

In preferred embodiments, the comparison can include: assigning an expected frequency from the class constant table to one or a plurality of the observed nearest neighbor pairs and determining how many of the observed nearest neighbor pairs fall above or below a predetermined value; determining the likelihood of occurrence, as predicted by the class constant table, for an observed nearest neighbor pair.; or determining if an observed nearest neighbor pair of a first and a second amino acid residue from the protein structure is predicted by the class constant table to occur at a predetermined frequency.

In preferred embodiments: the assignment of amino acids into a class is done by assigning a codon which encodes it into a class as a function of classifying triplets, e.g., the subject codon or a leading and following triplet of the subject codon, as a member of a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of triplets, wherein a binary choice parameter is a function of a message-level property of the nucleic acid sequence.

In preferred embodiments the method includes making a record of observed class constant nearest neighbors in the protein structure on a machine-readable medium.

In preferred embodiments: the method further includes determining if an observed nearest neighbor of the protein structure is that predicted, at a predetermined frequency, by the table, thereby evaluating the protein structure.

In preferred embodiments the method can be used to identify coding regions in a nucleic acid sequence. A coding region can be identified by comparing observed nearest neighbors in the protein structure with a class constant nearest neighbor table, the presence of observed pairs which correspond to predicted pairs in the table being predictive of a coding region. In a preferred embodiment, a codon in the coding region is changed so as to alter its encoded amino acid.

In preferred embodiments the method can identify structure or function critical residues, the occurrence of a nearest neighbor of low probability being predictive of a critical amino acid residue. In a preferred embodiment, a codon is changed so as to alter the amino acid encoded by the critical residue, a residue adjacent to the critical residue, or a residue which interacts with the critical residue.

In preferred embodiments: the protein structure is from a protein of known or unknown function.

In preferred embodiments: the protein structure is evaluated for the presence of a first nearest neighbor with a predicted occurrence below a predetermined value which is located in a run of residues, wherein at least 20, 40, 80, 90 or 95% of the residues in the run are members of nearest neighbors pairs having an expected frequency from the table of greater than a predetermined value, thereby identifying a critical residue. In a preferred embodiment, a codon is changed so as to alter the amino acid encoded by the critical residue, a residue adjacent to the critical residue, or a residue which interacts with the critical residue.

In preferred embodiments: the nearest neighbor includes or is adjacent to a critical residue.

In another aspect, the invention includes, a machine-readable medium on which is recorded a class-constant nearest neighbor table.

In another aspect, the invention features a method of evaluating a protein structure for resistance to change, e.g., evolutionary or mutational change. The method includes:

identifying regions of a protein which is encoded by runs of a single subcode, thereby identifying regions which have been resistant to change and which are therefor predicted to be functionally or structurally significant. E.g., the method can include

determining if the nucleic acid sequence which encodes the protein structure includes a run of triplets, e.g., a run at least 20, 40, 60, or 120 triplets in length (or e.g., 16, 32, 48, 64, 128, or 256 triplets in length), in which at least 20, 40, 60, 80, 90 or 95 %, or all, of the triplets in the run are from one class. Any of the ways of generating classes described herein can be used in this method.

In another aspect, the invention includes, a method of evaluating a protein structure for the presence of critical amino acid residues. The method includes:

identifying critical amino acid residues by identifying "minority codons" in runs encoded by codons of a single class or subcode, thereby identifying residues which have been resistant to change and which are therefor believed to be functionally important. Any of the ways of generating classes described herein can be used in this method.

In preferred embodiment: the evaluation comprises identifying a triplet from a first class in a s run of triplets of a second class, e.g., a run at least 20, 40, or 60 codons in length, in which at least 20, 40, 60, 80, 90 or 95 %, or all, of the codons are from the second class, thereby identifying the triplet of the first class as encoding a critical residue, e.g., a structure or function critical residue. In a preferred embodiment, a codon is changed so as to alter the amino acid encoded by the critical residue, a residue adjacent to the critical residue, or a residue which interacts with the critical residue.

In another aspect, the invention features, a method for evaluating a protein structure. The method includes:

providing a nucleic acid sequence which encodes the protein structure;

assorting bases of the nucleic acid sequence into subject triplets; and

assigning at least one of the subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets, the at least four classes of triplets being represented in at least a portion of the nucleic acid sequence in a ratio of about 3:5:3:5;

thereby evaluating the protein structure.

In preferred embodiments n is 1, 2, 3, or 4.

In preferred embodiments the method includes making a record, e.g., on a machine-readable medium, of the class assigned to one or more triplets.

In preferred embodiments, the classes can be generated by application of a binary choice parameter referred to herein.

In another aspect, the invention features, a method for identifying coding regions of a nucleic acid sequence, the method comprising:

providing the nucleic acid sequence;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets;

assigning the plurality of subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets A, B, C, and D;

determining whether the plurality of subject triplets are distributed into the at least four classes of triplets A:B:C:D in a ratio of about 3:5:3:5;

thereby identifying coding regions of the nucleic acid sequence.

In preferred embodiments n is 1, 2, 3, or 4.

In preferred embodiments (A+B)/(C+D) is about one.

In preferred embodiments (A+D)/(B+C) is about one.

In preferred embodiments the method includes making a record, e.g., on a machine-readable medium, of the class assigned to one or more triplets.

In preferred embodiments, the classes can be generated by application of a binary choice parameter referred to herein.

In another aspect, the invention features, a method for identifying a protein that includes a polypeptide portion which is structurally or functionally similar to all or a portion of a test protein, the method comprising:

providing a nucleic acid sequence which encodes all or a portion of the test protein;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets in a first reading frame;

assigning the plurality of subject triplets in the first reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a first binary choice alphabet of n degrees of freedom by applying n first binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

assorting bases of the at least a portion of the nucleic acid sequence into a plurality of subject triplets in a second reading frame;

assigning the plurality of subject triplets in the second reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a second binary choice alphabet of n degrees of freedom by applying n second binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5; and

identifying a protein which includes a polypeptide portion encoded by the plurality of triplets in the second reading frame;

thereby identifying a protein that includes a polypeptide portion which is structurally or functionally similar to all or a portion of the test protein.

In preferred embodiments each of the first and second binary choice alphabets, n is 1, 2, 3, or 4.n is two.

In preferred embodiments (A+B)/(C+D) is about one.

In preferred embodiments (A+D)/(B+C) is about one.

In preferred embodiments the first reading frame is frame 1 and the second reading frame is frame 2 or 3.

In preferred embodiments the method includes making a record, e.g., on a machine-readable medium, of the class assigned to one or more triplets.

In preferred embodiments, the classes can be generated by application of a binary choice parameter referred to herein.

In preferred embodiments the step of identifying a protein which includes a polypeptide portion encoded by the plurality of triplets in the second reading frame comprises reading all or a portion of a protein sequence from a database of protein sequences.

In another aspect, the invention features, a method for identifying a mutation-prone region of a nucleic acid sequence, e.g., a viral nucleic acid sequence. The method includes:

providing the nucleic acid sequence;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets in a first reading frame;

assigning the plurality of subject triplets in the first reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

assorting bases of the at least a portion of the nucleic acid sequence into a plurality of subject triplets in a second reading frame; and

assigning the plurality of subject triplets in the second reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

thereby identifying a mutation-prone region of the nucleic acid sequence.

In preferred embodiments the method includes making a record, e.g., on a machine-readable medium, of the class assigned to one or more triplets.

In preferred embodiments, the classes can be generated by application of a binary choice parameter referred to herein.

<u>Structural Binary Choice Parameters</u>

Structural binary choice parameters can be selected from a variety of physical or physico-chemical qualities related to the structure of the nucleic acid (polynucleotide) sequence, including the primary or secondary structure of the nucleic acid sequence, the

physical or chemical nature of the nucleotide bases, the physical or chemical nature of the codons, and the like. Thus, for example, properties related to the ability of a nucleic acid sequence to form secondary structure, e.g., by hybridization of subsequences of the nucleic acid sequence, can be selected as binary choice parameters. For example, the self-pairing of a nucleic acid sequence could be greater in, e.g., a highly UA (or GC)-rich region of the nucleic acid, while a nucleic acid which is not UA (or GC)-rich would be less prone to self-pairing.

Other exemplary binary choice parameters include the size of the nucleotide bases (e.g., pyrimidine vs. purine), H-bonding qualities due to H-bond donor or acceptor substituents (e.g., amino vs. keto-containing nucleotide bases), and the like.

Binary choice parameters can also be related to selected properties of codons, including the relative enthalpy of codon-anticodon interactions (which can include the relative enthalpy of the interaction of a codon with its anticodon plus the flanking complementary bases, e.g., the relative enthalpy of pentamers with their antiparallel complements; the ability of a codon to be "read" by a tRNA (which can be related to codon-anticodon interaction enthalpy, size, polarity, and the like), and other such codon-level parameters. However, a codon-level parameter is not a function of the amino acid encoded by the codon.

Compositional Binary Choice Parameters

Compositional binary choice parameters can be selected from observed frequencies of certain codon groups in one message reading frame and/or correlations among frequencies of particular codon groups in two different reading frames of the same message. Compositional choice parameters include those derived from enthalpic and statistical analysis of mRNA pentamers; compositional choice parameters also include any derived from energetic and statistical analysis of mRNA n-mers (i.e., $n > 3$), where such analyses can be shown to yield constant intra- and inter-frame frequencies of particular codon groups.

Application of Binary Choice Parameters

The application of a first binary choice parameter, e.g., with choices a and b, will structure the triplets into classes (or subcodes) a and b. The application of a second choice parameter, with choices c and d, will structure the triplets into classes ac, ad, bc, and bd. Application of a third binary choice parameter would structure triplets into 2^3 or eight subcodes. Thus, application of n binary choice parameters to the genetic code will result in the formation of a binary choice alphabet having 2^n classes or subcodes. It is

possible that some subcodes will be empty when the binary choice alphabet is applied to a given nucleic acid sequence.

The binary choice parameter can be applied directly to a subject triplet to assign triplets into a class. For example, the binary choice parameter can be based upon relative enthalpy of a codon-anticodon interaction (e.g., the codons are divided into group(s) of codons having high relative enthalpy and group(s) of codons having low relative enthalpy) and that parameter applied to a subject codon such as 234. A subject triplet can also be assigned a class by a method in which bases are not in the subject triplet, or which do not correspond exactly to the bases of the subject triple. E.g., the binary choice parameter can be applied to one or more base pairs which do not define the triplet. E.g, in evaluation of triplet 234, the binary choice parameter can be applied to triplet 123 and triplet 345, and the classes into which the triplets 123 and 345 fall can be used to assign a class or subcode to the triplet 234. In other words, the subcode of 234 can be a function of the application of the binary choice parameter to the triplets 123 and 345.

Frame Choice

Methods of the invention require the division of a sequence of bases into triplets. The simplest way is to consider a string of bases, 123456789, as triplets of 123 456 789. Mechanistically, this or any mode of division into triplets can be viewed as a process with two components, a "ratchet" or advance component and a "read" or selection component. As will be seen below, the ratchet component varies by the number of base pairs advanced after the determination of a triple.

The read component refers to the length in base pairs, of the segment of base pairs from which the triplet will be chosen.

The simplest system, that used by most evolutionarily current cellular mechanisms, is "ratchet three/read 3" (that is, the mRNA is advanced, or ratcheted, through the reading mechanism three bases at a time, and the message is read by the reading mechanism in groups of three bases (one codon). Other systems, however, are possible. Without being bound by theory, it is postulated that other systems may have existed in earlier stages in the evolution of the cellular protein translation machinery. In fact, examples of current frame-shift repressing tRNA's are known. Thus, possible alternate systems include "ratchet 3/read 5 on center" (in which the mRNA is ratcheted into the reading mechanism three bases at a time, and the reading mechanism reads the group of three bases at the center of a group of five bases in the reading mechanism). If a read value is more than 3 (e.g., in a "ratchet 3/read 5 on center" system), then additional choices are imposed: the triplet must be selected from the 3+N bases which

are read. Thus, a string 1 2 3 4 5 6 7 8 9 10 can be divided into the following triplets: 234 | 567 | 8910, which would be generated by reads 1*234*5 | 4*567*8 | 7*8910*11, wherein the italicized bases, the on center bases, are chosen.

For example, read-ratchet mechanisms or configurations can be divided into the following classes:

Class 1 : ratchet 3; read 3

Class 2 : ratchet 3; read 5 and select the center triplet, 1*234*5

Class 2a: ratchet 3; read 5 and select the leading triplet, *123*45

Class 2b: ratchet 3; read 5 and select the final triplet, 12*345*

Class 2c: ratchet 3; read 5 and read any triplet

Class 3 : ratchet 3; frameshift; read 5, any triplet

Class 2 approaches allow the assignment of a binary choice parameter to a codon 234 as a function of the binary choice parameter outcome for one or both of 123 and 345, e.g., 123 is classified as UG (k) or AC- (a) rich; and 345 is classified as UG (k) or AC-(a) rich, which gives the following possible classes for 234: kk, ka, aa, ak. If, for example, 123 is k, and 234 is a, then 234 is ka. Note that although only one binary choice is applied, there are 4 degrees of freedom with regard to 234, because the binary choice parameter is applied twice.

A binary choice parameter which divides triplets into classes on the basis of enthalpy, e.g., of the codon-anticodon interaction (e.g., into enthalpically strong and enthalpically weak classes) is particularly useful.

Read-ratchet configurations wherein the read value is greater than 3 make possible the context-sensitive (as opposed to context-free) assignment of triplets into classes by binary choice parameters, e.g., allow triplet 234 to be assigned a value which is a function of the binary choice parameter outcome of one and, more preferably both, of 123 and 345.

Binary Choice Alphabets

A binary choice alphabet can be constructed by selection of suitable pre-selected binary choice parameters. For example, binary choice parameters corresponding to enthalpy (of codon-anticodon interaction), size, polarity, charge, hydrophobicity, etc. can be selected and combined in any desired combination to arrive at a binary choice alphabet. Alternatively, a binary choice alphabet can be constructed by segregation of

codons into 2^n classes, without selecting the groups based on binary choice parameters. For example, a computer can rapidly segregate codons into randomly-selected classes to create a binary choice alphabet.

However the binary choice alphabet is constructed, it is generally preferable to validate the alphabet to ensure that the alphabet will be predictive of protein structure. It has been found that, in preferred embodiments, valid binary choice alphabets having 2^n classes will generally include at least four classes (A, B, C, D) for which the following relationships are true when triplets of a nucleic acid sequence are parsed with the binary choice alphabet: the ratio A:B:C:D is about 3:5:3:5 (e.g., from about 2:4:2:4 to about 7:11:7:11); the ratio (A+B)/(C+D) is about 1 (e.g., from about 0.9 to about 1.1); and the ratio (A+D)/(B+C) is about 1 (e.g., from about 0.9 to about 1.1). Thus, in preferred embodiments, application of a binary choice alphabet to a nucleic acid sequence which encodes a protein will yield at least four classes in which triplets are arrayed according to these ratios. It is therefore possible to validate a binary choice alphabet by searching for the appearance of the desired ratios. If the ratios are found, then the alphabet may have predictive value for protein structure evaluation. If the ratios are not found, the alphabet may not have such predictive value. It will be appreciated that the presence or absence of the ratios provides a useful "check" for a selected binary choice alphabet. In preferred embodiments, regions of triplets are checked for lengths which are multiples of 16 (e.g., 3+5+3+5=16), such as 16 triplets, 32 triplets, 64 triplets, and the like. Thus, in a preferred embodiment, groups of N x 16 sequential triplets (wherein N is an integer, e.g., between 1 and 16) are evaluated to determine whether the desired ratios are present.

Another means for validating a binary choice alphabet is by comparing the frequency of codon groups when the message is read in one reading frame (e.g., Frame 1) with the frequency of the same codon groups when the message is read in another reading frame (e.g., Frame 3). It has also been found that valid binary choice alphabets will generally include at least four classes A, B, C, D such that the frequency of codons A, B, C, D varies systematically from one frame to another, e.g., from Frame 2 to Frame 1, Frame 2 to Frame 3, and/or Frame 1 to Frame 3. It is therefore possible to further validate a binary choice alphabet by searching for a systematic inter-frame variation in the frequencies of codon groups defined by the alphabet. If such systematic variation is found, the alphabet may have predictive value. One of ordinary skill in the art, in light of the teachings herein, will be able to select useful binary choice alphabets according to these criteria using no more than routine experimentation.

Examples

Example 1: Generation of a predictive amino acid alphabet based on binary choices which are a function of enthalpy of codon-anticodon interaction

This example provides a predictive four letter amino acid alphabet ($4a^3$) for the representation of protein primary structures (s^{1s}) from the energetic properties of mRNA molecules, i.e., the translation of mRNAs. While not wishing to be bound by theory, the basis for deriving an amino acid alphabet from codon-anticodon interactions can be rationalized as follows: if the genetic code was not "frozen" prior to the onset of translation and the evolution of protein primary structure, then the evolutionary trajectory of this code may have been one factor which determined important properties of protein primary structure. Energetics of codon-anticodon interactions may have been relevant to the evolution of the genetic code before ribosomes existed, when these interactions occurred in an aqueous medium.

The configuration of the reading frame may also provide a basis for deriving an amino acid alphabet. Figure 1 schematically depicts two alternative reading frames for a nucleic acid sequence, each reading frame defining an energetic packet or triplet; each nucleic acid base of the message is represented by a black square. Again, while not being bound by theory, evolution may have favored systems which would allow slippage from frame 1 to frame 2. This would impose entropic requirements on the code. It is noted that this in system, which permits "slippage", energy packaging may be analogous to human linguistic systems which permit slippage and routines for assigning syllabic stress (f^0) and consequent systematic recasting of signifying sound tokens. For comparison, refer to Grimm's Law and Verner's Law for Indo-European.

It is shown herein that the energies of codon-anticodon interactions pattern systematically, that this pattern implicitly defines a particular amino acid alphabet, and that this amino acid alphabet characterizes protein primary structure predictively, i.e., provides insight into protein secondary and tertiary structure.

Table IA shows the Ornstein-Fresco ΔH values for the 10 possible base pair overlaps. Using those ΔH values, , average ΔH values were calculated for all 64 possible codon-anticodon interactions for all possible mRNA pentamers (or "five-envelopes") with codons as the center three bases. The average ΔH values shown in Tables IB and IC assume that there is no wobble pairing of codon and anticodon. The average ΔH values in Tables 1B and C were calculated according to the following formula: for any pentamer ABCDE: ΔH is calculated for B, C, D according to the formula: (ΔH(AB) +ΔH (BC) + ΔH (CD) + ΔH (BE)) / 4. The 64 codon triplets are shown in the first column of

Table IB. Values for a codon in each of all possible five-envelopes are shown in each row. For example, in the case of UUU, the enthalpy value for a UUU codon preceded by a U and followed by a U is 2.80. The enthalpic value when UUU is preceded by a U and is followed by a C is 2.45. The average value for a codon in all possible "five-envelopes" is given in the penultimate column on the right side of the table. For the UUU codon, the average for all possible 5 envelopes is 2.43. That average is calculated for all codons in Table IB. The final column (far right) of Table IB provides the average enthalpic value for all codons having a common leading doublet. For example, all codons which begin with the doublet UU have an average enthalpic value of 2.11.

Table IC shows the values from the penultimate column of Table IB. Note that the values in Table IC hover around four values, 0.6, 1.2, 1.8, and 2.4. It can also be seen, as indicated in the caption of Table IC, that for any given doublet XX, the average enthalpic value for the codons XXU and XXA is about 0.6 higher than the average value for the codons XXC and XXG.

The energetic pattern evident in Table IC manifests itself in mRNAs. Table IIA shows 16 enthalpically defined codon groups (separated by dashed lines) produced by ranking the codons according to the interaction ΔH of the leading doublet, that is, the first two base pairs of the codon, and by the codon interaction enthalpy value from Table 1C. In Table IIA the first column shows all codons. The second column identifies the first doublet in the third bases of the codon. The third column provides the ΔH of the first doublet, the fourth column provides the main codon ΔH over all 16 possible pentameric envelopes (as set out in Table IB, penultimate column) and the fifth column provides a letter for a group designation. The horizontal divisions segregate the first doublets according to the eight energy levels shown in Table IA. Each of the groups thus formed by horizontal division is further subdivided on the basis of the average value for the codon for each of the 5 possible envelopes for Table IB and by which of the 4 energy levels identified in Table IC it falls into. Table IIB is analogous except that the first binary choice applied is the ΔH for the second or final doublet of the codon.

Table IIC shows the frequency of Table IIA, or leading, codon groups and of Table IIB, or following codon groups in a test mRNA database.

The leading or L codon groups of Table IIC correspond to frame 1 of the mRNA and the final or F codon groups in Table IIC correspond to frame 3 of the mRNA. The middle column of Table IIC shows the difference in frequency between the L groups and the F groups shown in the first and last columns of Table IIC. It can be seen that the differences are very small, which may be a consequence of an original evolutionary pentameric energy packaging scheme.

One possible explanation for this conserved "epiphenomenon" is that the present day "ratchet 3/read 3" translation system evolved from a "ratchet 3/read 5 on center" primordial translation system. Present day "frame shift suppressor" tRNAs with anticodon loops greater than 3, are possibly mutant analogs of ancestor tRNAs which regularly read pentamers. According to this view, as ratchet 3/read 3 translation systems evolved from ratchet 3/read 5 ancestral translation systems, mRNAs would have had to be repackaged in one of two alternative reading frames different from the original reading frame. For example, an original evolutionary ratchet 3/read 5 on center system would read pentamer 12345 as 234. This corresponds to present day frame 2. However, a ratchet 3/read 3 translation system reading from that same pentamer 12345, would read 123, corresponding to present day reading frame 1, or else it would read 345, corresponding to present day reading frame 3. It is believed that the prevalence of the "weak" bases U and A at the 5' ends of the anticodon loops of tRNA pentamers would favor repackaging of codons into present day frame 1 rather than into present day frame 3.

If such an evolution from a ratchet 3/read 5 on center to a ratchet 3/read 3 translation system occurred, the resulting frameshift from reading frame 2 to reading frame 1 would have the potential to cause disastrous changes in protein structure as the alternate reading frame was read. There are at least two ways in which catastrophic mutations could be avoided. First, if the pentamer packets of the earliest mRNAs were read "loosely" by the earliest tRNA anticodon, that is, if early tRNAs could read either 123, 234, or 345 out of each pentamer, then the loose reading would result in evolutionary pressure to select mRNAs containing packets which would not introduce harmful amino acids into protein primary structures when the packets were read differently, e.g., when the packets were read in frame 1 rather than in frame 2. Second, if the mRNAs were so selected from the start, then a systemic frameshift would not necessarily introduce harmful amino acids into protein primary structures in numbers sufficient to damage structure and/or function of the protein, and in fact might permit the introduction of novel amino acid sequences with beneficial effects on protein secondary and tertiary structures.

This suggests that if a systemic frameshift occurred, some codon distributions would have remained essentially unchanged ("constant" codons) while other codon distributions would have changed ("wild card"), which could have a beneficial effect on protein structure. In this case, the evolutionary distinction between "wild card" and "constant" codons might classify amino acids in such a way as to enable the construction of a predictive amino acid alphabet. Accordingly, a binary choice alphabet was created

in which the "constant" vs. "wildcard" distinction was one binary choice parameter (Figure 2).

Table IIIA shows possible enthalpic groups of leading and final triplets in mRNA pentamers with the 64 codons as centers. An example is shown in Figure 2, in which the codon UUA is the center triple. The first column of Figure 2 shows the four possible leading L triplets together with the classification group from Table IIA in the second column. The fourth column of Figure 11 shows the classification group of the final (F) triplets shown in the last column of Figure 11.

As shown in Table IIIA, doublets can be classified as "constant codon doublets" or "wild card codon doublets". A constant codon doublet is a doublet XX of a codon XXY or XXR (Y and R stand for a pyrimidine base or a purine base respectively), in which XX is UU, CC, GG, or AA, for which codon, as shown in Table IIIA, the leading (NXX) and final (XYN or XRN) triplets of all possible pentamers (N is any base), belong to the same enthalpic groups of Tables IIA and IIB. For example, for the codon UUA (boxed line at upper left of Table IIIA), the four possible leading triplets (NUU) all belong to the groups Z and W. The four possible final triplets (UAN) also all belong to the groups Z, W, and X. Because U is a pyrimidine (Y) and A is a purine (R), UUA is a constant codon doublet of class YXR. A "wild card codon doublet", in contrast, shows an alternation between enthalpic groups of Tables IIA and IIB as the leading and final triplets are analyzed over all pentamers. For example, for the codon UUU (top line at upper left of Table IIIA), the four possible leading triplets (NUU) belong to the groups Z, W and X, as noted above. The four possible final triplets (UUN) belong to the groups Z, V, Y, and U, differing from the leading triplets. Because U is a pyrimidine (Y), UUU is a constant codon doublet of class YXY.

The distinction between constant codon doublets and wild card codon doublets can be used to construct a four letter amino acid alphabet. As shown in Figure 3, the 64 codons can be divided into four groups: constant Y, X, R, doublets, constant R, X, Y doublets, and wild card Y, X, Y, doublets, and wild card R, X, R doublets.

As shown in Figure 4, a test mRNA database was analyzed to determine the frequencies of the four codon groups in the four letter amino acid alphabet of Figure 3. The mRNA database was read in both frame 1 and frame 2. As can be seen from Figure 4, shifting from reading in frame 2 to reading in frame 1 results in the interchange of frequencies of p and s.

Example 2: Determination of Secondary and Tertiary Protein Structural Features Correlated With Message Segments Evaluated With a Binary Choice Alphabet

A binary choice alphabet of Example 1 (s, p, d, t) was used to evaluate protein structures as follows:

Test mRNA sequences were analyzed from a database of mRNAs (e.g., from GenBank). Note that in GenBank, uracil (U) is stored as "T"; this convention will be used throughout this example. Each sequence was then analyzed in reading frame 2 using the following mapping:

ATT/ATC/GTT/GTC=A
ACT/ACC/GCT/GCC=B
AAT/AAC/GAT/GAC=C
AGT/AGC/GGT/GGC=D
TTA/TTG/CTA/CTG=E
TCA/TCG/CCA/CCG=F
TAA/TAG/CAA/CAG=G
TGA/TGG/CGA/CGG=H
TTT/TTC/CTT/CTC=I
TCT/TCC/CCT/CCC=J
TAT/TAC/CAT/CAC=K
TGT/TGC/CGT/CGC=L
ATA/ATG/GTA/GTG=M
ACA/ACG/GCA/GCG=N
AAA/AAG/GAA/GAG=O
AGA/AGG/GGA/GGG=P

One binary choice parameter was whether the leading base of the triplet was purine (A or G; groups A-D and M-P) or pyrimidine (T or C; groups E-L). The other binary choice parameter was the "wildcard" vs. "constant" distinction discussed in Example 1, infra. It should be noted that this parameter also corresponds to a binary choice between "symmetrical" (YXY and RXR) codons vs. "non-symmetrical" (YXR and RXY) codons (in which Y and R are pyrimidine and purine as defined above).

The mapped string from reading frame 2 was then converted to the binary choice alphabet (s, p, d, t) according to the following scheme: ABCDEFGHIJKLMNOP=ssssppppddddtttt. The result is a binary choice alphabet of degree 2, dividing the genetic code into 4 classes (denoted s, p, d, t), as shown in Figure 3.

The mapped string was then evaluated, over a moving window of 16 triplets (16 letters in the spdt alphabet), to determine regions in which the s:p:d:t ratio was about 3:5:3:5 in reading frame 2 (that is, s >= 2, p >= 4, d >= 2, t >= 4). When such a region was found, the mRNA sequence was translated to an amino acid sequence in frame 1 for that region of the mRNA (i.e., by reading the message resulting from adding a base at the beginning and eliminating a base at the end of the message segment). Our protein database (described *infra*) was then searched for proteins which included the amino acid sequence encoded by the resulting Frame 1 amino acid sequence. When a single protein was found to have two separate and distinct regions with even low homology to the derived Frame 1 amino acid sequence, the two regions were often found to have similar, or virtually identical, secondary and tertiary structural features. When two different proteins were found, which each manifested one or more regions with even low homology to the derived Frame 1 amino acid sequence, these regions were often found to have very similar secondary and tertiary structural features.

Example 3: Starting from a Known Protein Structure

Binary choice alphabets (s, p, d, t) were used to evaluate protein structures as follows:

Test mRNA sequences were read from a database of mRNAs (e.g., from GenBank). Each sequence was then read in reading frame 1 and in reading frame 2 using the mapping described in Example 2 for the 16-letter alphabet A-P.

The mapped string from reading frame 1 was then converted to a binary choice alphabet (s, p, d, t) according to the following scheme:

ABCDEFGHIJKLMNOP=ppppsssstttdddd

The mapped string from reading frame 2 was then converted to a binary choice alphabet (s, p, d, t) according to the following scheme:

ABCDEFGHIJKLMNOP=sssspppddddtttt

The mapped strings were then evaluated, over a moving window of 16 triplets (16 letters in the spdt alphabet), to determine regions in which the s:p:d:t ratio was about 3:5:3:5 in both frame 1 and frame 2. When such a region was found, the mRNA sequence was translated to an amino acid sequence in both frame 1 and frame 2 for that region of the mRNA. Our protein database was then searched for proteins which contain the amino acid sequence encoded by the translated region of Frame 2. The database of protein messages contained messages for three hundred proteins, those proteins being sixty to six thousand amino acids in length. The proteins included proteins with roles in protein synthesis, nucleic acid synthesis, protein or nucleic acid degradation, various "house-keeping" enzymes, and some immunoglobulins. When a protein containing the

sequence was found, the structural similarity (e.g., the tertiary structure) of that portion of the protein was compared to the structure of the protein encoded by the test mRNA sequence.

It was found that for several test mRNA sequences, many of those portions of the identified proteins were structurally very similar to the comparable portions of the protein encoded by the test mRNA sequence. For example, a helix-strand transition in the protein encoded by the test mRNA sequence was structurally similar to a helix-strand transition of a protein located in the protein database according to methods of the invention. Application of the methods of the invention (e.g., the methods of Example 2 and Example 3) to a variety of test sequences identified structural similarity in at least one protein of the our protein database for other structural motifs such as sheets, helix entry, helix exit, Pro-His-Pro turns, and the like.

Example 4

The function of introns (e.g., non-coding DNA sequences in genomic DNA) is generally not well understood. Methods of the invention provide knowledge which is useful for investigating intron function. The methods of the invention can include searching nucleic acid databases (e.g., of genomic DNA) for regions of nucleic acid which do not code for protein in the present-day reading frame (i.e., frame 1), but which could code for protein in an alternate reading frame (e.g., frame 2 or frame 3). Such a presently non-coding region (i.e., an intron) could correspond to a region of a nucleic acid which was a coding region prior to a frameshift. Such formerly-coding regions could encode alternate structures (i.e., protein regions which differ from the modern protein regions) which preserve the function of the protein.

Thus, a nucleic acid which represents both coding and non-coding regions can be analyzed in both frames 1 and 2, as described *supra* for Examples 2 and 3. Where a non-coding region, such as an intron, is found in which the s:p:d:t ratio is about 3:5:3:5 in frame 2, that region may correspond to a region of the nucleic acid which coded for protein structure prior to a shift in reading frame.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims. The contents of all references cited herein are hereby incorporated by reference.

What is claimed is:

1. A method of evaluating protein structure comprising:

providing a nucleic acid sequence which encodes the protein structure;

assorting bases of the nucleic acid sequence into subject triplets; and

assigning one or a plurality of subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying triplets, of the nucleic acid sequence as members of a class of a binary choice alphabet of n degrees of freedom, and wherein the classes can be generated by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein a binary choice parameter is a function of a message-level property of the nucleic acid sequence,

thereby evaluating the protein structure.

2. The method of claim 1, further comprising making a record, on a machine readable medium of the class assigned to one or more triplets.

3. The method of claim 1, wherein triplets are assigned to a first and a second class:

the first class having the property that a message made of triplets drawn exclusively from the first class is less likely to form secondary (intrachain) structure than is a message which is made of triplets from both the first class and the second class of triplets, and

the second class having the property that a message made of triplets drawn exclusively from the second class is less likely to form secondary (intrachain) structure than is a message which is made of triplets from both the first class and the second class of triplets.

4. The method of claim 1, wherein the message-level property is: a function of the UA content of a subject triplet; a function of the GC content of a subject triplet; a function of the size or molecular weight of a triplet; a function of whether the triplet is keto rich or amino rich; a function of whether the triplet is purine rich or pyrimidine rich; or a function of a the enthalpy of the interaction between the triplet and a fully or partially complementary nucleic acid.

5. The method of claim 1, wherein a subject triplet 456 of a nucleic acid sequence of bases 123 456 789 is assigned into a class as a function of:

(1) performing one or more of (i), (ii), and (iii)

(i) applying a binary choice parameter to a leading triplet of 456, e.g., to one or more of triplet 123, 234, or 345, to yield a leading value;

(ii) applying a binary choice parameter to 456, to provide a center value;

(iii) applying a binary choice parameter to a following triplet of 456, e.g., to one or more of triplet 567, 678, or 789, to yield a following value;

(2) assigning one or a plurality of subject triplets 345 into a class based on the values determined in one or more of (1), (3) and (3),

thereby assigning one or a plurality of subject triplets into classes.

6. A class-constant table of nearest neighbor relationships for amino acid residues which provides, for each of a plurality of class constant nearest neighbors, a frequency of occurrence which is a function of the occurrence of the class constant nearest neighbor pair in a collection of at least 10 proteins.

7. A method of evaluating a protein structure comprising:

providing a class-constant table of nearest neighbor relationships for amino acid residues;

providing a nucleic acid which encodes a protein structure; and

comparing one or a plurality of the observed nearest neighbor pairs in the protein structure with the frequencies provided by the class constant table, thereby evaluating the protein structure.

8. The method of claim 7 wherein the comparison can include: assigning an expected frequency from the class constant table to one or a plurality of the observed nearest neighbor pairs and determining how many of the observed nearest neighbor pairs fall above or below a predetermined value; determining the likelihood of occurrence, as predicted by the class constant table, for an observed nearest neighbor pair; or determining if an observed nearest neighbor pair of a first and a second amino acid residue from the protein structure is predicted by the class constant table to occur at a predetermined frequency.

9. The method of claim 7, further comprising making a record of observed class constant nearest neighbors in the protein structure on a machine-readable medium.

10. A machine-readable medium on which is recorded a class-constant nearest neighbor table.

11. A method of evaluating a protein structure for resistance to change, e.g., evolutionary or mutational change comprising:

identifying regions of a protein which is encoded by runs of a single subcode, thereby identifying regions which have been resistant to change and which are therefor predicted to be functionally or structurally significant.

12. The method of claim 11, wherein the method includes determining if the nucleic acid sequence which encodes the protein structure includes a run of triplets at least 40 triplets in length, in which at least 90% of the triplets in the run are from one class.

13. A method of evaluating a protein structure for the presence of critical amino acid residues comprising:

identifying critical amino acid residues by identifying minority codons in runs encoded by codons of a single class or subcode, thereby identifying residues which have been resistant to change and which are therefor believed to be functionally important.

14. The method of claim 13, wherein the evaluation comprises identifying a triplet from a first class in a run of triplets of a second class at least 40 codons in length, in which at least 40% of the codons are from the second class, thereby identifying the triplet of the first class as encoding a critical residue.

15. A method for evaluating a protein structure comprising:

providing a nucleic acid sequence which encodes the protein structure;

assorting bases of the nucleic acid sequence into subject triplets; and

assigning at least one of the subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets, the at least four classes of triplets being represented in at least a portion of the nucleic acid sequence in a ratio of about 3:5:3:5;

thereby evaluating the protein structure.

16. The method of claim 15, wherein the method includes making a record on a machine-readable medium of the class assigned to one or more triplets.

17. A method for identifying coding regions of a nucleic acid sequence, the method comprising:

providing the nucleic acid sequence;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets;

assigning the plurality of subject triplets to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets A, B, C, and D;

determining whether the plurality of subject triplets are distributed into the at least four classes of triplets A:B:C:D in a ratio of about 3:5:3:5;

thereby identifying coding regions of the nucleic acid sequence.

19. The method of claim 17, wherein the method includes making a record on a machine-readable medium, of the class assigned to one or more triplets.

20. A method for identifying a protein that includes a polypeptide portion which is structurally or functionally similar to all or a portion of a test protein, the method comprising:

providing a nucleic acid sequence which encodes all or a portion of the test protein;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets in a first reading frame;

assigning the plurality of subject triplets in the first reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a first binary choice alphabet of n degrees of freedom by applying n first binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

assorting bases of the at least a portion of the nucleic acid sequence into a plurality of subject triplets in a second reading frame;

assigning the plurality of subject triplets in the second reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a second binary choice alphabet of n degrees of freedom by applying n second binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5; and

identifying a protein which includes a polypeptide portion encoded by the plurality of triplets in the second reading frame;

thereby identifying a protein that includes a polypeptide portion which is structurally or functionally similar to all or a portion of the test protein.

21. A method for identifying a mutation-prone region of a viral nucleic acid sequence comprising:

providing the nucleic acid sequence;

assorting bases of at least a portion of the nucleic acid sequence into a plurality of subject triplets in a first reading frame;

assigning the plurality of subject triplets in the first reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

assorting bases of the at least a portion of the nucleic acid sequence into a plurality of subject triplets in a second reading frame; and

assigning the plurality of subject triplets in the second reading frame to one of a plurality of classes, wherein the assignment is a function of classifying the subject triplets of the nucleic acid sequence under a binary choice alphabet of n degrees of freedom by applying n binary choice parameters to a triplet to yield at least 2^n classes of subject triplets, wherein the assignment provides at least four classes of triplets distributed in a ratio of about 3:5:3:5;

thereby identifying a mutation-prone region of the nucleic acid sequence.

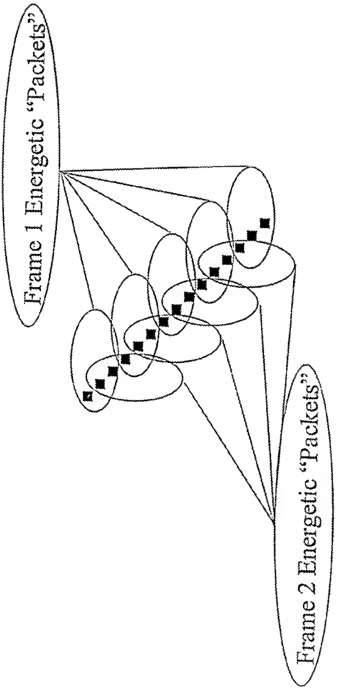

FIGURE 1

"Wildcards" and "Constants": The Key to a Predictive $4A^{3}$'

Table IIIA: Possible Enthalpic Groups of Leading and Final Triples in mRNA pentamers with the 64 Codons as Centers

Example: Pentamers with Codon "uua" as Center Triple

Possible Leading "L" Triples	Group of "L" Triple	Center and (Codon)	Group of "F" Triple	Possible Leading "F" Triples
uuu	Z	uua	Z	uau
cuu	W	uua	W	uag
auu	Z	uua	Z	uaa
guu	X	uua	X	uac

FIGURE 2

A $4A^3$ Denoting Four Codon Groups Defined Using the "Constant:Wildcard" Distinction

Constant

p = Constant YXR Doublets				s = Constant RXY Doublets			
L:uua	S:uca	*:uaa	*:uga	I:auu	T:acu	N:aau	S:agu
L:uug	S:ucg	*:uag	W:ugg	I:auc	T:acc	N:aac	S:agc
L:cua	P:cca	Q:caa	R:cga	V:guu	A:gcu	D:gau	R:ggu
L:cug	P:ccg	Q:cag	R:cgg	V:guc	A:gcc	D:gac	R:ggc

YXX ← → RXX

F:uuu	S:ucu	Y:uau	*:ugu	I:aua	T:aca	N:aau	R:aga
F:uuc	S:ucc	Y:uac	W:ugc	I:aug	T:acg	N:aac	R:agg
L:cuu	P:ccu	Q:cau	R:cgu	V:gua	A:gca	D:gaa	R:gga
L:cuc	P:ccv	Q:cac	R:cgc	V:gug	A:gcg	D:gag	R:ggg
d = Wildcard YXY Doublets				t = Wildcard RXR Doublets			

Wildcard

FIGURE 3

Frequencies of Four Codon Groups in $4A^3$ in Frame +1 and Frame +2 of Test mRNA Database

Frame +1

$f_p = 51902$	$f_s = 31905$
$f_d = 32114$	$f_t = 52650$

Note:

$(f_s + f_d)/(f_p + f_t) = 0.6123$,

or 0.0057 from τ

Frame +2

$f_p = 28130$	$f_s = 56097$
$f_d = 36047$	$f_t = 48925$

Note:

f_s *and* f_p *are interchanged with respect to frame +1;* f_d *and* f_t *are not*

FIGURE 4

Table IA: RNA Doublet Relative ΔH Enthalpies (from Ornstein,Fresco)

au/ua	2.86
aa/uu	2.80
uu/aa	2.80
ua/au	2.07
gu/ca	1.91
ac/ug	1.91
cu/ag	1.52
ag/uc	1.52
uc/ag	1.41
ga/cu	1.41
ug/ac	1.16
ca/gu	1.16
gc/cg	0.95
cc/gg	0.27
gg/cc	0.27
cg/gc	0.00

FIGURE 5

Table IB

ΔH_{mf} Values for Each of the 64 Codon-Anticodon Interactions in All Possible Pentamers

Center	16 Possible Surrounding 5' and 3' Bases																	Avg	Avg
codon	u-u	u-c	c-u	c-c	a-a	a-g	g-a	g-g	u-g	g-u	a-c	c-a	u-a	a-u	c-g	g-c		(Row)	(Qrtet)
uuu	2.80	2.45	2.48	2.13	2.63	2.41	2.40	2.17	2.39	2.58	2.47	2.30	2.62	2.82	2.07	2.23		2.43	2.11
uua	2.63	2.40	2.31	2.08	2.63	2.31	2.40	2.08	2.30	2.41	2.41	2.30	2.62	2.65	1.98	2.17		2.36	
uuc	2.13	1.82	1.81	1.50	2.06	1.77	1.82	1.53	1.75	1.91	1.84	1.72	2.04	2.15	1.43	1.60		1.81	
uug	2.17	1.93	1.85	1.61	2.06	1.77	1.82	1.54	1.76	1.95	1.94	1.72	2.04	2.18	1.44	1.71		1.84	
cuu	2.13	1.79	1.85	1.50	2.08	1.85	1.84	1.61	1.72	2.02	1.91	1.67	1.95	2.26	1.44	1.67		1.83	1.51
cua	1.97	1.73	1.68	1.44	2.08	1.76	1.84	1.52	1.63	1.85	1.85	1.67	1.95	2.09	1.35	1.61		1.75	
cuc	1.47	1.15	1.18	0.87	1.50	1.21	1.26	0.97	1.09	1.35	1.28	1.09	1.38	1.59	0.80	1.04		1.20	
cug	1.50	1.28	1.22	0.98	1.50	1.22	1.26	0.98	1.09	1.39	1.39	1.09	1.38	1.63	0.81	1.15		1.24	
auu	2.63	2.29	2.41	2.06	2.63	2.41	2.29	2.08	2.22	2.47	2.47	2.22	2.45	2.82	2.00	2.12		2.35	2.02
aua	2.47	2.23	2.24	2.00	2.63	2.31	2.29	1.97	2.13	2.30	2.41	2.22	2.45	2.65	1.90	2.06		2.27	
auc	1.97	1.65	1.74	1.43	2.05	1.77	1.71	1.42	1.59	1.80	1.84	1.65	1.88	2.15	1.36	1.49		1.72	
aug	2.00	1.76	1.77	1.53	2.06	1.77	1.71	1.43	1.59	1.84	1.94	1.65	1.88	2.18	1.36	1.60		1.75	
guu	2.17	1.82	1.88	1.53	2.08	1.85	1.76	1.54	1.76	1.95	1.91	1.70	1.99	2.26	1.47	1.60		1.83	1.50
gua	2.00	1.76	1.71	1.47	2.08	1.76	1.76	1.44	1.67	1.78	1.85	1.70	1.99	2.09	1.38	1.54		1.75	
guc	1.50	1.19	1.21	0.90	1.50	1.21	1.19	0.90	1.12	1.28	1.28	1.12	1.41	1.59	0.83	0.97		1.20	
gug	1.54	1.30	1.25	1.01	1.50	1.22	1.19	0.90	1.13	1.31	1.39	1.12	1.41	1.63	0.84	1.07		1.24	
ucu	2.13	1.79	1.81	1.47	1.97	1.74	1.73	1.50	1.72	1.91	1.80	1.63	1.95	2.15	1.40	1.55		1.77	1.48
uca	2.06	1.82	1.74	1.50	2.06	1.74	1.82	1.50	1.72	1.84	1.84	1.72	2.04	2.07	1.40	1.60		1.78	
ucc	1.50	1.19	1.18	0.87	1.43	1.14	1.19	0.90	1.12	1.28	1.20	1.09	1.41	1.52	0.80	0.97		1.17	
ucg	1.53	1.29	1.21	0.97	1.42	1.14	1.18	0.90	1.12	1.31	1.31	1.09	1.41	1.55	0.80	1.07		1.21	
ccu	1.50	1.15	1.22	0.87	1.44	1.22	1.20	0.98	1.09	1.39	1.28	1.03	1.32	1.63	0.81	1.04		1.20	0.91
cca	1.43	1.19	1.14	0.90	1.54	1.22	1.30	0.98	1.09	1.31	1.31	1.13	1.41	1.55	0.81	1.07		1.21	
ccc	0.87	0.56	0.58	0.27	0.90	0.61	0.66	0.37	0.49	0.75	0.68	0.49	0.78	0.99	0.20	0.44		0.60	
ccg	0.90	0.66	0.61	0.37	0.90	0.61	0.66	0.37	0.49	0.78	0.78	0.49	0.77	1.02	0.20	0.54		0.63	
acu	2.08	1.73	1.85	1.50	2.08	1.85	1.73	1.58	1.67	1.91	1.91	1.67	1.89	2.26	1.44	1.56		1.79	1.50
aca	2.00	1.76	1.77	1.54	2.17	1.85	1.82	1.58	1.67	1.84	1.95	1.76	1.99	2.18	1.44	1.60		1.80	
acc	1.44	1.13	1.22	0.90	1.54	1.25	1.19	0.90	1.06	1.28	1.31	1.13	1.35	1.63	0.84	0.97		1.20	
acg	1.47	1.23	1.25	1.01	1.53	1.25	1.18	0.90	1.06	1.31	1.42	1.12	1.35	1.66	0.84	1.07		1.23	

FIGURE 6

Table IB (Continued)

ΔH_{nl} Values for Each of the 64 Codon-Anticodon Interactions in All Possible Pentamers

Center codon	16 Possible Surrounding 5' and 3' Bases																	Avg	Avg
codon	u-u	u-c	c-u	c-c	a-a	a-g	g-a	g-g	u-g	g-u	a-c	c-a	u-a	a-u	c-g	g-c		(Row)	(Qrtet)
gcu	1.61	1.25	1.32	0.97	1.52	1.29	1.20	0.98	1.20	1.39	1.35	1.14	1.43	1.70	0.91	1.04		1.27	0.98
gca	1.53	1.30	1.24	1.01	1.61	1.29	1.38	0.98	1.28	1.31	1.39	1.23	1.52	1.62	0.91	1.07		1.28	
gcc	0.98	0.66	0.69	0.37	0.98	0.69	0.66	0.37	0.53	0.75	0.75	0.60	0.89	1.07	0.31	0.44		0.68	
gcg	1.01	0.77	0.72	0.48	0.97	0.69	0.66	0.37	0.60	0.78	0.86	0.59	0.88	1.10	0.31	0.54		0.71	
uau	2.63	2.29	2.31	1.97	2.47	2.24	2.23	2.00	2.22	2.41	2.30	2.13	2.45	2.65	1.90	2.06		2.27	2.03
uaa	2.63	2.40	2.31	2.08	2.63	2.31	2.40	2.08	2.30	2.41	2.41	2.30	2.62	2.65	1.98	2.17		2.36	
uac	2.08	1.78	1.76	1.44	2.00	1.71	1.76	1.47	1.70	1.85	1.78	1.67	1.99	2.09	1.38	1.54		1.75	
uag	2.08	1.84	1.78	1.52	1.97	1.68	1.73	1.44	1.67	1.85	1.85	1.63	1.95	2.09	1.35	1.61		1.75	
cau	2.06	1.71	1.77	1.43	2.00	1.77	1.76	1.53	1.65	1.94	1.84	1.59	1.88	2.18	1.39	1.60		1.75	1.52
caa	2.05	1.82	1.77	1.54	2.17	1.85	1.93	1.61	1.72	1.94	1.95	1.76	2.04	2.18	1.44	1.71		1.84	
cac	1.50	1.19	1.22	0.90	1.54	1.25	1.30	1.01	1.12	1.39	1.31	1.13	1.41	1.63	0.84	1.07		1.24	
cag	1.50	1.26	1.22	0.98	1.50	1.22	1.26	0.98	1.09	1.39	1.39	1.09	1.38	1.63	0.81	1.15		1.24	
aau	2.63	2.29	2.41	2.06	2.53	2.41	2.29	2.06	2.22	2.47	2.47	2.22	2.45	2.82	2.00	2.12		2.35	2.11
aaa	2.63	2.40	2.41	2.17	2.80	2.48	2.45	2.13	2.30	2.47	2.58	2.39	2.62	2.82	2.07	2.23		2.43	
aac	2.08	1.76	1.85	1.54	2.17	1.88	1.82	1.53	1.70	1.91	1.85	1.76	1.99	2.26	1.47	1.60		1.83	
aag	2.08	1.84	1.85	1.61	2.13	1.85	1.79	1.50	1.67	1.91	2.02	1.72	1.95	2.25	1.44	1.67		1.83	
gau	2.06	1.71	1.77	1.42	1.97	1.74	1.65	1.43	1.65	1.84	1.80	1.59	1.88	2.15	1.36	1.49		1.72	1.48
gaa	2.06	1.82	1.77	1.53	2.13	1.81	1.82	1.50	1.72	1.84	1.91	1.75	2.04	2.15	1.43	1.60		1.81	
gac	1.50	1.19	1.21	0.90	1.50	1.21	1.19	0.90	1.12	1.28	1.28	1.12	1.41	1.59	0.83	0.97		1.20	
gag	1.50	1.26	1.21	0.97	1.47	1.18	1.15	0.87	1.09	1.28	1.35	1.09	1.38	1.59	0.80	1.04		1.20	
ugu	2.17	1.82	1.65	1.50	2.00	1.77	1.76	1.54	1.76	1.95	1.84	1.67	1.99	2.18	1.44	1.60		1.80	1.52
uga	2.06	1.82	1.74	1.50	2.06	1.74	1.82	1.50	1.72	1.84	1.84	1.72	2.04	2.07	1.40	1.60		1.78	
ugc	1.61	1.30	1.29	0.98	1.53	1.24	1.30	1.01	1.23	1.39	1.31	1.20	1.52	1.62	0.91	1.07		1.28	
ugg	1.54	1.30	1.22	0.98	1.43	1.14	1.19	0.90	1.13	1.31	1.31	1.09	1.41	1.55	0.81	1.07		1.21	
cgu	1.53	1.18	1.25	0.90	1.47	1.25	1.23	1.01	1.12	1.42	1.31	1.08	1.35	1.66	0.84	1.07		1.23	0.94
cga	1.42	1.18	1.14	0.90	1.53	1.21	1.29	0.97	1.09	1.31	1.31	1.12	1.41	1.55	0.80	1.07		1.21	
cgc	0.97	0.66	0.69	0.37	1.01	0.72	0.77	0.48	0.59	0.86	0.78	0.60	0.88	1.10	0.31	0.54		0.71	
cgg	0.90	0.66	0.61	0.37	0.93	0.61	0.66	0.37	0.49	0.78	0.78	0.43	0.77	1.02	0.20	0.54		0.63	

TABLE 6 (continued)

Table IB (Continued)

ΔH_M Values for Each of the 64 Codon-Anticodon Interactions in All Possible Pentamers

Center	16 Possible Surrounding 5' and 3' Residues																	Avg	Avg
codon	u-u	u-c	c-u	c-c	a-a	a-g	g-a	g-g	u-g	g-u	a-c	c-a	u-a	a-u	c-g	g-c		(Row)	(Qrtet)
agu	2.08	1.73	1.85	1.50	2.08	1.85	1.73	1.50	1.67	1.91	1.91	1.67	1.69	2.25	1.44	1.56		1.79	1.51
aga	1.97	1.73	1.74	1.50	2.13	1.81	1.79	1.47	1.63	1.80	1.91	1.72	1.95	2.15	1.40	1.56		1.77	
agc	1.52	1.20	1.29	0.98	1.61	1.32	1.25	0.97	1.14	1.35	1.39	1.20	1.43	1.70	0.90	1.04		1.27	
agg	1.44	1.20	1.22	0.98	1.50	1.22	1.15	0.67	1.03	1.28	1.39	1.09	1.32	1.63	0.81	1.04		1.20	
ggu	1.54	1.19	1.25	0.90	1.44	1.22	1.13	0.90	1.13	1.31	1.28	1.06	1.35	1.63	0.84	0.97		1.20	0.91
gga	1.43	1.19	1.14	0.90	1.50	1.18	1.19	0.87	1.09	1.20	1.28	1.12	1.41	1.52	0.80	0.97		1.17	
ggc	0.98	0.66	0.69	0.37	0.98	0.69	0.66	0.37	0.60	0.75	0.75	0.60	0.39	1.07	0.31	0.44		0.68	
ggg	0.90	0.66	0.61	0.37	0.87	0.58	0.56	0.27	0.49	0.68	0.75	0.49	0.78	0.99	0.20	0.44		0.60	
Column avg:																			
	1.78	1.50	1.50	1.22	1.78	1.50	1.50	1.22	1.40	1.60	1.60	1.40	1.68	1.89	1.12	1.32			

FIGURE 6 (continued)

Table IC: Codon Square Showing ΔH Values Per Codon Averaged Over The 16 Possible "5-Envelopes"

Note: 1) All Values Hover Around 0.6, 1.2., 1.8, or 2.4
2) $(\Delta H(XXc) + \Delta H(XXg))/2 - (\Delta H(XXu) + \Delta H(XXa))/2 = @\text{-}0.6$
(for XX = same leading doublet)

PHE uuu 2.43	SER ucu 1.77	TYR uau 2.27	CYS ugu 1.80
LEU uua 2.36	SER uca 1.78	*** uaa 2.36	*** uga 1.78
PHE uuc 1.81	SER ucc 1.17	TYR uac 1.75	CYS ugc 1.28
LEU uug 1.84	SER ucg 1.21	*** uag 1.75	TRP ugg 1.21
LEU cuu 1.83	PRO ccu 1.20	HIS cau 1.75	ARG cgu 1.23
LEU cua 1.75	PRO cca 1.21	GLN caa 1.84	ARG cga 1.21
LEU cuc 1.20	PRO ccc 0.60	HIS cac 1.24	ARG cga 0.71
LEU cug 1.24	PRO ccg 0.63	GLN cag 1.24	ARG cgg 0.63
ILE auu 2.35	THR acu 1.79	ASN aau 2.35	SER agu 1.79
ILE aua 2.27	THR aca 1.80	LYS aaa 2.43	ARG aga 1.77
ILE auc 1.72	THR acc 1.20	ASN aac 1.83	SER agc 1.27
MET aug 1.75	THR acg 1.23	LYS aag 1.83	ARG agg 1.20
VAL guu 1.83	ALA gcu 1.27	ASP gau 1.72	GLY ggu 1.20
VAL gua 1.75	ALA gca 1.28	GLU gaa 1.81	GLY gga 1.17
VAL guc 1.20	ALA gcc 0.68	ASP gaq 1.20	GLY ggc 0.68
VAL gug 1.24	ALA gcg 0.71	GLU gag 1.20	GLY ggg 0.60

FIGURE 7

Table IIA: Codon Table Doubly Ranked by ΔH of Leading Doublet and Mean Codon ΔH over 16 Possible 5-Envelopes

Note: W = u,a and S = c,g

Codon	-Bases- 1-2	3	1-2 ΔH	Mean Codon ΔH Over 16 Possible 5-Envelopes	Group Designation
auu	a u	u	2.86	2.35	Z_L
uuu	u u	u	2.80	2.43	
aau	a a	u	2.80	2.35	
aua	a u	a	2.86	2.27	
uua	u u	a	2.80	2.36	
aaa	a a	a	2.80	2.43	
auc	a u	c	2.86	1.72	R_L
uuc	u u	c	2.80	1.81	
aac	a a	c	2.80	1.83	
aug	a u	g	2.86	1.75	
uug	u u	g	2.80	1.84	
aag	a a	g	2.80	1.83	
uau	u a	u	2.07	2.27	Y_L
uaa	u a	a	2.07	2.36	
uac	u a	c	2.07	1.75	Q_L
uag	u a	g	2.07	1.75	
guu	g u	u	1.91	1.83	X_L
acu	a c	u	1.91	1.79	
gua	g u	a	1.91	1.75	
aca	a c	a	1.91	1.80	
guc	g u	c	1.91	1.20	P_L
acc	a c	c	1.91	1.20	
gug	g u	g	1.91	1.24	
acg	a c	g	1.91	1.23	
cuu	c u	u	1.52	1.83	W_L
agu	a g	u	1.52	1.79	
cua	c u	a	1.52	1.75	
aga	a g	a	1.52	1.77	
cuc	c u	c	1.52	1.20	O_L
agc	a g	c	1.52	1.27	
cug	c u	g	1.52	1.24	
agg	a g	g	1.52	1.20	

FIGURE 8

Table IIA (Continued)

Note: W = u,a and S = c,g

Codon	-Bases- 1-2	3	1-2 ΔH	Mean Codon ΔH Over 16 Possible 5-Envelopes	Letter Designation
ucu	u c	u	1.41	1.77	V_L
gau	g a	u	1.41	1.72	
uca	u c	a	1.41	1.78	
gaa	g a	a	1.41	1.81	
ucc	u c	c	1.41	1.17	N_L
gac	g a	c	1.41	1.20	
ucg	u c	g	1.41	1.21	
gag	g a	g	1.41	1.20	
ugu	u g	u	1.16	1.80	U_L
cau	c a	u	1.16	1.75	
uga	u g	a	1.16	1.78	
caa	c a	a	1.16	1.84	
ugc	u g	c	1.16	1.28	M_L
cac	c a	c	1.16	1.24	
ugg	u g	g	1.16	1.21	
cag	c a	g	1.16	1.24	
gcu	g c	u	0.95	1.27	T_L
gca	g c	a	0.95	1.28	
gcc	g c	c	0.95	0.68	L_L
gcg	g c	g	0.95	0.71	
ccu	c c	u	0.27	1.20	S_L
ggu	g g	u	0.27	1.20	
cgu	c g	u	0.00	1.23	
cca	c c	a	0.27	1.21	
gga	g g	a	0.27	1.17	
cga	c g	a	0.00	1.21	
ccc	c c	c	0.27	0.60	K_L
ggc	g g	c	0.27	0.68	
cgc	c g	c	0.00	0.71	
ccg	c c	g	0.27	0.63	
ggg	g g	g	0.27	0.60	
cgg	c g	g	0.00	0.63	

FIGURE 8 (continued)

Table IIB: Codon Table Doubly Ranked by ΔH of Final Doublet and Mean Codon ΔH over 16 Possible 5-Envelopes

Note: W = u,a and S = c,g

Codon	-Bases- 1	2-3	2-3 ΔH	Mean Codon ΔH Over 16 Possible 5-Envelopes	Letter Designation
uau	u	a u	2.86	2.27	Z_r
uuu	u	u u	2.80	2.43	
uaa	u	a a	2.80	2.36	
aau	a	a u	2.86	2.35	
auu	a	u u	2.80	2.35	
aaa	a	a a	2.80	2.43	
cau	c	a u	2.86	1.75	R_r
cuu	c	u u	2.80	1.83	
caa	c	a a	2.80	1.84	
gau	g	a u	2.86	1.72	
guu	g	u u	2.80	1.83	
gaa	g	a a	2.80	1.81	
uua	u	u a	2.07	2.36	Y_r
aua	a	u a	2.07	2.27	
cua	c	u a	2.07	1.75	Q_r
gua	g	u a	2.07	1.75	
ugu	u	g u	1.91	1.80	X_r
uac	u	a c	1.91	1.75	
agu	a	g u	1.91	1.79	
aac	a	a c	1.91	1.83	
cgu	c	g u	1.91	1.23	P_r
cac	c	a c	1.91	1.24	
ggu	g	g u	1.91	1.20	
gac	g	a c	1.91	1.20	
ucu	u	c u	1.52	1.77	W_r
uag	u	a g	1.52	1.75	
acu	a	c u	1.52	1.79	
aag	a	a g	1.52	1.83	
ccu	c	c u	1.52	1.20	O_r
cag	c	a g	1.52	1.24	
gcu	g	c u	1.52	1.27	
gag	g	a g	1.52	1.20	

FIGURE 9

Table IIB (Continued)

Note: W = u,a and S = c,g

Codon	-Bases- 1	-Bases- 2-3	2-3 ΔH	Mean Codon ΔH Over 16 Possible 5-Envelopes	Letter Designation
uuc	u	u c	1.41	1.81	V_r
uga	u	g a	1.41	1.78	
auc	a	u c	1.41	1.72	
aga	a	g a	1.41	1.77	
cuc	c	u c	1.41	1.20	N_r
cga	c	g a	1.41	1.21	
guc	g	u c	1.41	1.20	
gga	g	g a	1.41	1.17	
uug	u	u g	1.16	1.84	U_r
uca	u	c a	1.16	1.78	
aug	a	u g	1.16	1.75	
aca	a	c a	1.16	1.80	
cug	c	u g	1.16	1.24	M_r
cca	c	c a	1.16	1.21	
gug	g	u g	1.16	1.24	
gca	g	c a	1.16	1.28	
ugc	u	g c	0.95	1.28	T_r
agc	a	g c	0.95	1.27	
cgc	c	g c	0.95	0.71	L_r
ggc	g	g c	0.95	0.68	
ucc	u	c c	0.27	1.17	S_r
ugg	u	g g	0.27	1.21	
ucg	u	c g	0.00	1.21	
acc	a	c c	0.27	1.20	
agg	a	g g	0.27	1.20	
acg	a	c g	0.00	1.23	
ccc	c	c c	0.27	0.60	K_r
cgg	c	g g	0.27	0.63	
ccg	c	c g	0.00	0.63	
gcc	g	c c	0.27	0.68	
ggg	g	g g	0.27	0.60	
gcg	g	c g	0.00	0.71	

FIGURE 9 (continued)

Table IIC

Comparative Frequencies of "L" Codon Groups
in Frame 0 (123...) and "F" Codon Groups in Frame +2 (345...)
(Protein mRNA Sample 1)

N = 169199

N = 168631

---- "L" Groups ----			L - F	---- "F" Groups ----		
Group	#	%	Δ%	%	#	Group
Z_L	20692	12.23	0.51	11.72	19766	Z_F
R_L	22334	13.20	-0.42	13.62	22968	R_F
Y_L	3091	1.71	0.03	1.68	2835	Y_F
Q_L	2965	2.07	-0.15	1.91	3220	Q_F
X_L	9911	5.86	-0.23	6.09	10273	X_F
P_L	11417	6.75	0.22	6.53	11004	P_F
W_L	7026	4.15	-0.22	4.37	7367	W_F
O_L	10966	6.48	0.20	6.28	10583	O_F
V_L	14880	8.79	0.12	8.67	14622	V_F
N_L	12384	7.32	-0.19	7.51	12663	N_F
U_L	5941	3.51	-0.47	3.98	6719	U_F
M_L	9331	5.51	0.45	5.06	8530	M_F
T_L	6091	3.60	-0.15	3.75	6327	T_F
L_L	6355	3.76	0.15	3.61	6082	L_F
S_L	12682	7.50	0.25	7.25	12226	S_F
K_L	13033	7.70	-0.27	7.97	13446	K_F

FIGURE 10

Table IIIA:
Enthalpic Groups of Leading and Final Triples in mRNA Pentamers Containing the 64 Codons as Center

Enthalpic Group When Lead Triple Is: u c a g	End Triple Is: u c a g	Enthalpic Group When Lead Triple Is: u c a g	End Triple Is: u c a g	Enthalpic Group When Lead Triple Is: u c a g	End Triple Is: u c a g	Enthalpic Group When Lead Triple Is: u c a g	End Triple Is: u c a g
Z W Z X uu	uu Z W Y U	R O R P uc	cu R N Q M	Z W Z X ua	au Z V Y U	R O R P ug	gu R N Q M
Z W Z X uu	uc W S U S	R O R P uc	cc O K M K	Z W Z X ua	ac W S U S	R O R P ug	gc O K M K
Z W Z X uu	ua Z X Z W	R O R P uc	ca R P R O	Z W Z X ua	aa Z X Z W	R O R P ug	ga R P R O
Z W Z X uu	ug X T V S	R O R P uc	cg P L N K	Z W Z X ua	ag X T V S	R O R P ug	gg P L N K
V S X T cu	uu Z V Y U	N K P L cc	cu R N Q M	V S X T ca	au Z V Y U	N K P L cg	gu R N Q M
W S X T cu	uc W S U S	N K P L cc	cc O K M K	V S X T ca	ac W S U S	N K P L cg	gc O K M K
V S X T cu	ua Z X Z W	N K P L cc	ca R P R O	V S X T ca	aa Z X Z W	N K P L cg	ga R P R O
V S X T cu	ug X T V S	N K P L cc	cg P L N K	V S X T ca	ag X T V S	N K P L cg	gg P L N K
Y U Z V au	uu Z V Y U	Q M R N ac	cu R N Q M	Y U Z V aa	au Z V Y U	Q M R N ag	gu R N Q M
Y U Z V au	uc W S U S	Q M R N ac	cc O K M K	Y U Z V aa	ac W S U S	Q M R N ag	gc O K M K
Y U Z V au	ua Z X Z W	Q M R N ac	ca R P R O	Y U Z V aa	aa Z X Z W	Q M R N ag	ga R P R O
Y U Z V au	ug X T V S	Q M R N ac	cg P L N K	Y U Z V aa	ag X T V S	Q M R N ag	gg P L N K
U S W S gu	uu Z V Y U	M K O K gc	cu R N Q M	U S W S ga	au Z V Y U	M K O K gg	gu R N Q M
U S W S gu	uc W S U S	M K O K gc	cc O K M K	U S W S ga	ac W S U S	M K O K gg	gc O K M K
U S W S gu	ua Z X Z W	M K O K gc	ca R P R O	U S W S ga	aa Z X Z W	M K O K gg	ga R P R O
U S W S gu	ug X T V S	M K O K gc	cg P L N K	U S W S ga	ag X T V S	M K O K gg	gg P L N K

FIGURE 11

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 97/19673

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C07K1/00 G06F17/50

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C07K G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BIOLOGICAL ABSTRACTS, vol. 98, no. 3, 1994 Philadelphia, PA, US; abstract no. 34267, XP002056708 see abstract & D HALITSKY: "A geometric model for codon recognition logic" MATHEMATICAL BIOSCIENCES, vol. 121, no. 2, 1994, pages 227-234, --- -/--	1-21

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

1

Date of the actual completion of the international search	Date of mailing of the international search report
23 February 1998	19. 03. 1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax. (+31-70) 340-3016	Masturzo, P

International Application No
PCT/US 97/19673

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	R L ORNSTEIN &J R FRESCO: "Correlation of crystallographically determined and computationally predicted hydrogen-bonded pairing configurations of nucleic acid bases" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA., vol. 80, no. 17, September 1983, WASHINGTON US, pages 5171-5174, XP002056707 see the whole document -----	1-21

1

INTERNATIONAL SEARCH REPORT

Int. ational application No
PCT/US 97/19673

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. [X] Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

see FURTHER INFORMATION sheet PCT/ISA/210

2. [ ] Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3. [ ] Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6 4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. [ ] As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. [ ] As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. [ ] As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. [ ] No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

[ ] The additional search fees were accompanied by the applicant's protest.

[ ] No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Remark : Although claims 6 and 10 are directed to a representation of information on a carrier, the search has been carried out and based on the molecular struucture represented by this information (Art. 17, Rule 39. 1 PCT).